Anomaly Detection

Dataset $\{x^{(1)}, x^{(2)}, \dots, x^{(m)}\}$

Density estimation.

$p(x_{test}) < \varepsilon \rightarrow$ flag anomaly.

$\geq \varepsilon \rightarrow$ ok.

how anamaly ?

Ex. Fraud detection.

- $x^{(i)}$ = features of user i's activities
- model $p(x)$ from data.
- identify unusual users by checking $p(x) < \varepsilon$.

Ex. Manufacturing

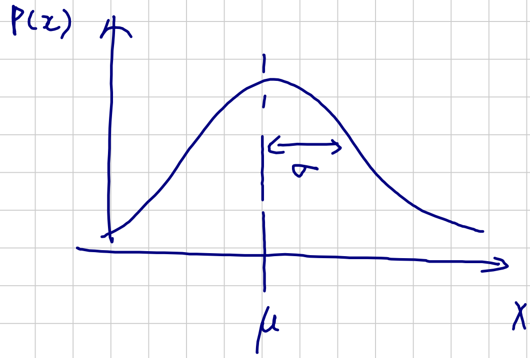
Ex. Monitoring computers in a data centre.

- $x^{(i)}$ = features of machine i.
- x_1 = memory use, x_2 = number of disk accesses / sec.
- x_3 = CPU load, x_4 = CPU load / network traffic.

Gaussian Distribution

$x \in \mathbb{R}$, $x \sim \mathcal{N}(\mu, \sigma^2)$ "normal"

μ: mean, σ^2: variance.

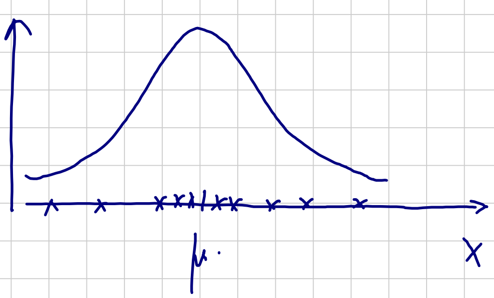


$$p(x; \mu, \sigma^2) \triangleq \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Parameter Estimation

Dataset $\{x^{(1)}, x^{(2)}, \dots, x^{(m)}\}$, $x^{(i)} \in \mathbb{R}$.

If we expect. $x^{(i)} \sim N(\mu, \sigma^2)$



$$\hat{\mu} = \frac{1}{m} \cdot \sum_{i=1}^{m} x^{(i)}$$

$$\sigma^2 = \frac{1}{m} \sum_{i=1}^{m} [x^{(i)} - \mu]^2$$

$m-1$. (sometimes)

Algorithm

Training set. $\{x^{(1)}, \dots, x^{(m)}\}$. Each example is $x \in \mathbb{R}^n$.

$$p(x) = \prod_{j=1}^{n} p(x_j; \mu_j, \sigma_j^2)$$

$x_j \sim N(\mu_j, \sigma_j^2)$

1. Choose features x_i that you think might be indicative of anomalous examples.
2. Fit parameters $\mu_1 \dots \mu_n$. $\sigma_1^2, \dots \sigma_n^2$ $(j = 1 \dots n)$

$$\mu_j = \frac{1}{m} \sum_{i=1}^{m} x_j^{(i)} \quad ; \quad \sigma_j^2 = \frac{1}{m} \sum_{i=1}^{m} (x_j^{(i)} - \mu_j)^2$$

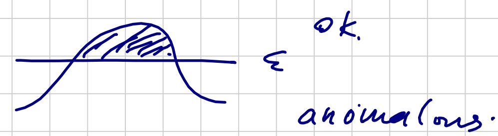
$$\underline{\mu} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix} = \frac{1}{m} \sum_{i=1}^{m} x^{(i)}$$

3. Given a new example x

$$p(x) = \prod_{j=1}^{n} p(x_j; \mu_j, \sigma_j^2)$$

$$= \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi}\,\sigma_j} \exp\left(-\frac{(x_j-\mu_j)^2}{2\sigma_j^2}\right) \overset{?}{<} \varepsilon \rightarrow \text{anamoly}$$

"Assuming a Gaussian distribution, determine anamoly based on a threshold ε"



Developing and Evaluating an Anomaly Detection

- real number evaluation.

Algorithm evaluation

- fit model $p(x)$ on training set. $\{x^{(1)}, \dots x^{(m)}\}$
- On a CV/ test x, predict. $y = \begin{cases} 1, & p(x) < \varepsilon \ \text{(anomaly)} \\ 0, & p(x) > \varepsilon \ \text{(normal)} \end{cases}$

↳ skewed data set. (should true/positive, precision / recall, F1 score)

Anomaly Detection v.s. Supervised Learning

Examples

$1 <$ $\underline{0}$ (negative examples) → fit $p(x)$

$\underline{1}$, $\underline{0}$

Many diff. types of anomalies

Enough positive examples for algorithm to get a sense of what positive examples are likely.

Choosing what Features to use

- Non-guassian $\longrightarrow$ $\log(x + \underbrace{c}_{\text{constant}})$, $x^{\frac{1}{a}}$ . ...

  (transformation)

- error analysis for anomaly detection
  - Want: $p(x)$ large for normal examples x

    $p(x)$ small for anomalous examples x.

  - common problem: $p(x)$ is comparable (say, both large) for normal and anomalous examples.

Multivariate Guassian.

$x \in \mathbb{R}^n$. Don't model $p(x_1), p(x_2), \dots$, etc. separately.

Model $p(x)$ all in one-go.

Parameters: $\mu \in \mathbb{R}^n$, $\Sigma \in \mathbb{R}^{n \times n}$. [covariance matrix].

$$p(x;\mu,\Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}} \cdot |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)$$

$\hookrightarrow$ determinant.

$\Sigma = \begin{pmatrix} 0.6 & 0 \\ 0 & 0.6 \end{pmatrix}$ $\longrightarrow$ variance shrinked $\rightarrow$ narrower distribution

$\Sigma = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$ flatter.

Anomaly Detection with the multivariate Guassian.

1) Fit model $p(x)$ by setting

$$\mu = \frac{1}{m}, \quad \Sigma = \frac{1}{m} \sum_{i=1}^{m} (x^{(i)} - \mu) \cdot (x^{(i)} - \mu)^T.$$

2) Given a new example x, compute

$$p(x) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} \exp\cdot\left(-\frac{1}{2} (x-\mu)^T \cdot \Sigma^{-1} (x-\mu) \right).$$

anomaly $\iff$ $p(x) < \varepsilon$.

- Original model $p(x) = \prod p\cdot(x_i, \mu_i, \sigma_i^2)$ $\leftrightarrow$ $p\cdot(x; \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} \exp\cdot\left(-\frac{1}{2}\cdot(x-\mu)^T \cdot \Sigma^{-1} \cdot (x-\mu)\right)$

  where. $\Sigma = \begin{pmatrix} \sigma & & \emptyset \\ & \ddots & \\ \emptyset & & \sigma \end{pmatrix}$

  special case.

- When to use?

original model.

$\prod p(x; \mu, \sigma)$

- Manually create features to capture anomalies where x_1, x_2 take unusual combinations of values. $x_3 = \frac{x_1}{x_2}$
- m can $< n$.

multivariate Guassian

$$p = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} \exp\cdot\left(-\frac{1}{2}(x-\mu)^T \cdot \Sigma^{-1} (x-\mu)\right)$$

- Automatically captures correlations between features.
- Computationally expensive.
- Σ invertible $\iff$ $m > n$. ($m \geq 10n$ in practice)

  ↳ redundant features when n ↑